BIG DATA FINAL PROJECT

ML STREAMING WITH SPARK
DATASET: SENTIMENT ANALYSIS

Team Members:
PES1UG19CS156 – Gagan G R
PES1UG19CS234 – Kshitij Prit Gopali
PES1UG19CS293 – Navya Eedula
PES1UG19CS433 – Sarthak Deva

- The current dataset consists of real-life tweets which are binary classified into positive (4) ad negative (0) tweets.
- An 80:20 split of train: validation was performed on the train.csv in order to train the model.

Preprocessing:
We used the following techniques to preprocess the data –
1. Remove stop words
2. Stemming
3. Remove punctuation marks
4. Tokenizer

Vectoriser:
HashTF

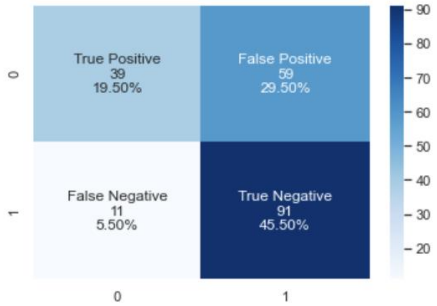
Models:

1. Stochastic Gradient Descent:
   Used to find the minimum loss function and the best fit between the predicted and target values.
2. Passive-Aggressive Classifier:
   It is an incremental model that is passive towards correctly classified instances but aggressive towards misclassified instances.
3. Bernoulli Naïve Bayes:
   Performs Naïve Bayes classification with binary variables. The model penalizes the least frequently occurring words.
4. K-means Clustering:
   Two clusters of positive and negative tweets are created for various batch sizes based on similarity.

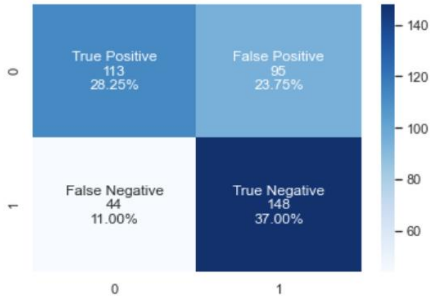
Plots:

1. Bernoulli Naïve Bayes:

a) Confusion matrix – Batch size – 1000

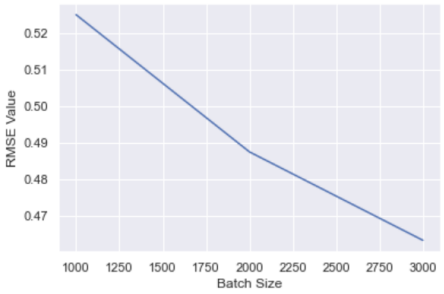


b) Confusion matrix – Batch size - 2000

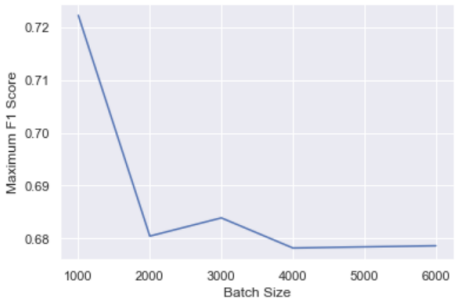


c) Batch size VS RMSE

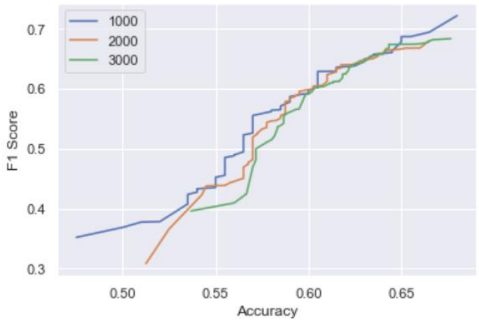
d) Batch size VS Best F1 Score

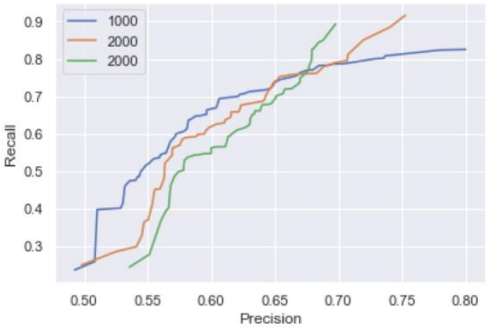


e) F1 Score VS Accuracy



f) Recall VS Precision

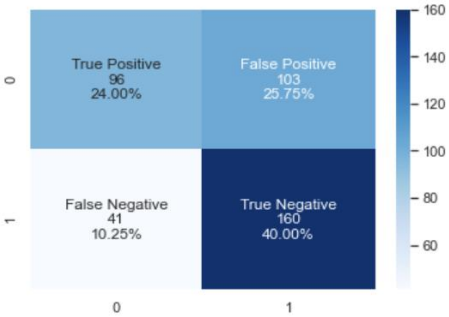




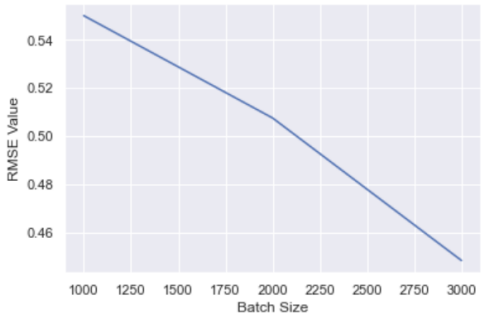
2. <u>Stochastic Gradient Descent:</u>

b) Confusion matrix – Batch size – 1000

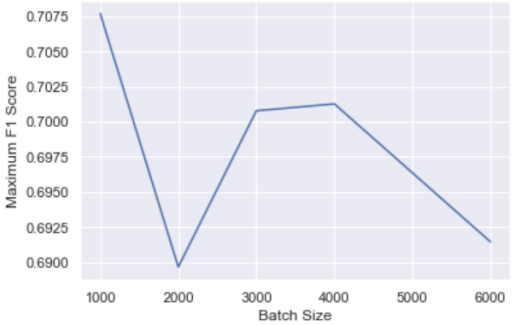
b) Confusion matrix – Batch size - 2000



d) Batch size VS RMSE

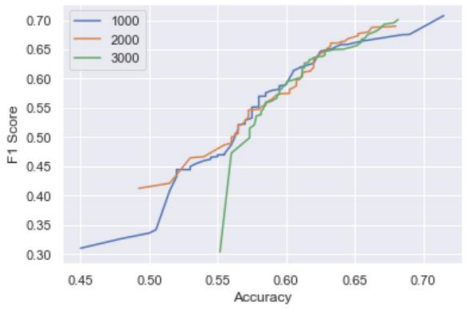


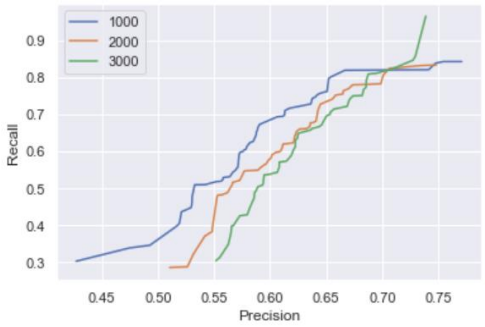
d) Batch size VS Best F1 Score



f) F1 Score VS Accuracy

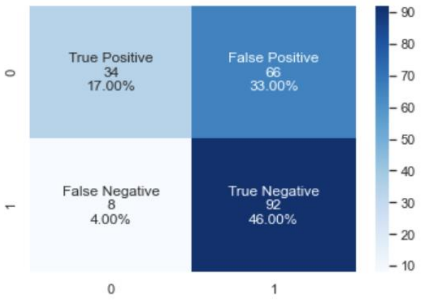
f) Recall VS Precision



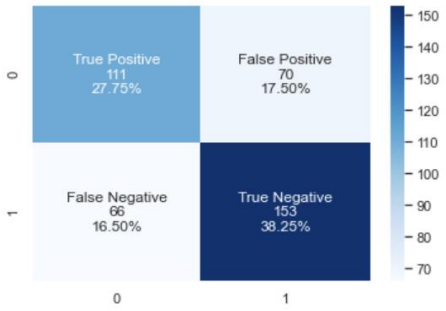


3. <u>Passive Aggressive Classifier:</u>

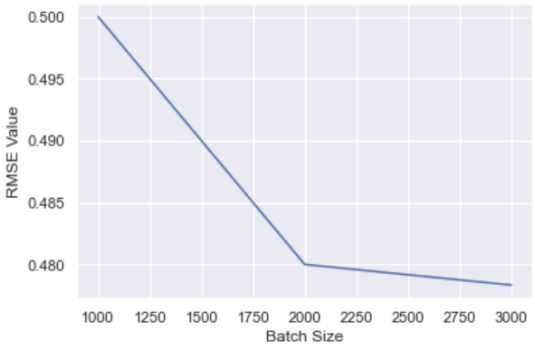
a) Confusion matrix – Batch size – 1000



b) Confusion matrix – Batch size - 2000



b) Batch size VS RMSE



d) Batch size VS Best F1 Score